Why study Statistics ??

- Data is everywhere
- Statistical techniques are used to make many decisions that affect our lives
  - When do you generally leave for office.
  - How much kitchen essentials do you purchase every cycle
- No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively
- Applications in Finance, marketing, Sales , Supply chain and Operations, IT . It encompasses virtually every domain

Statistics

To understand God's thoughts we must study statistics, for these are the measure of His purpose.

— Florence Nightingale

While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will be up to, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.

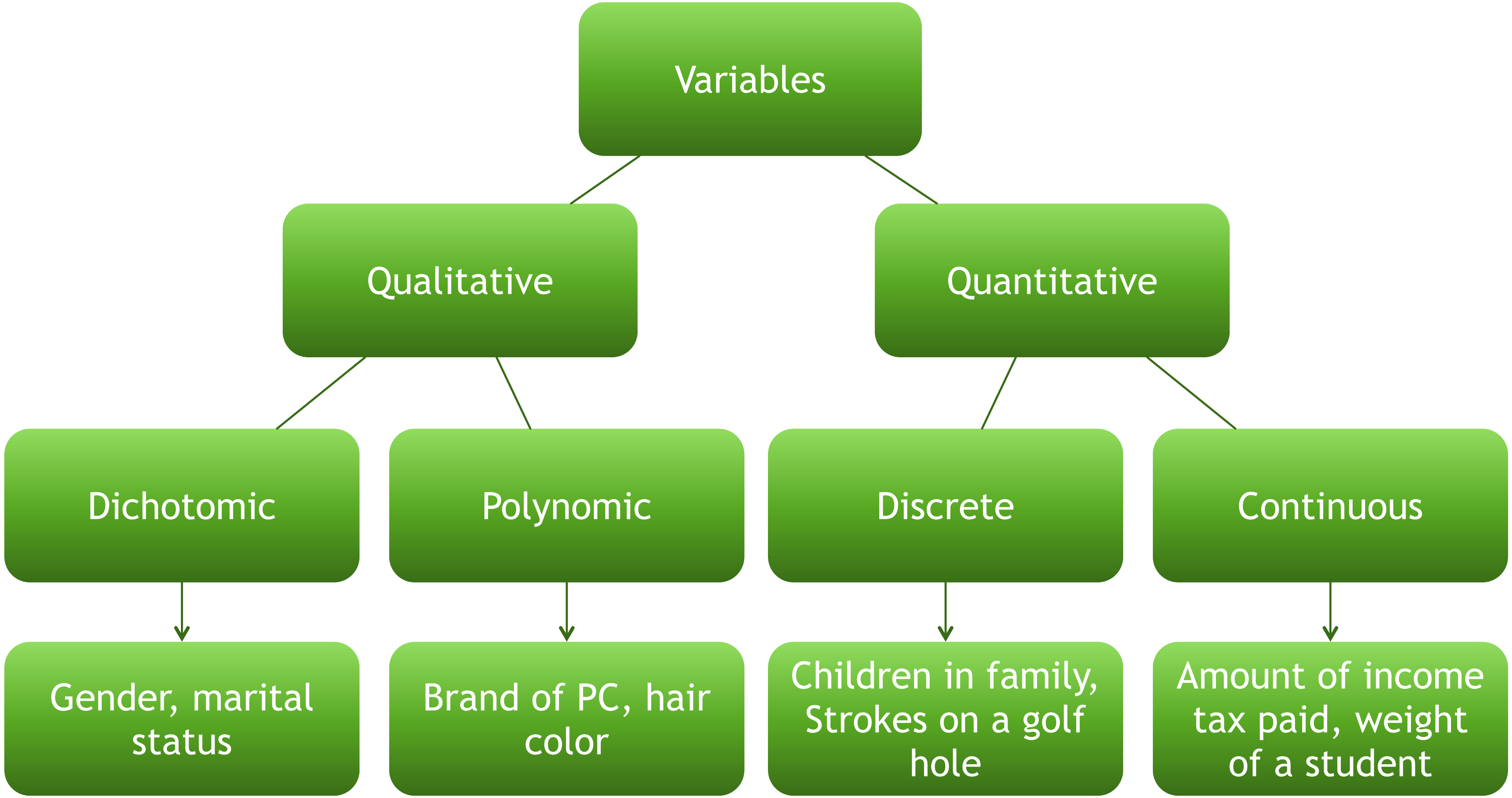
— Arthur Conan Doyle

Data from a statistical perspective

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the entire research project.
- Statistical data are usually obtained by counting or measuring items.
  - **Primary data** are collected specifically for the analysis desired
  - **Secondary data** have already been compiled and are available for statistical analysis
- A **variable** is an item of interest that can take on many different numerical values.
- A **constant** has a fixed numerical value.

Most data can be put into the following categories:

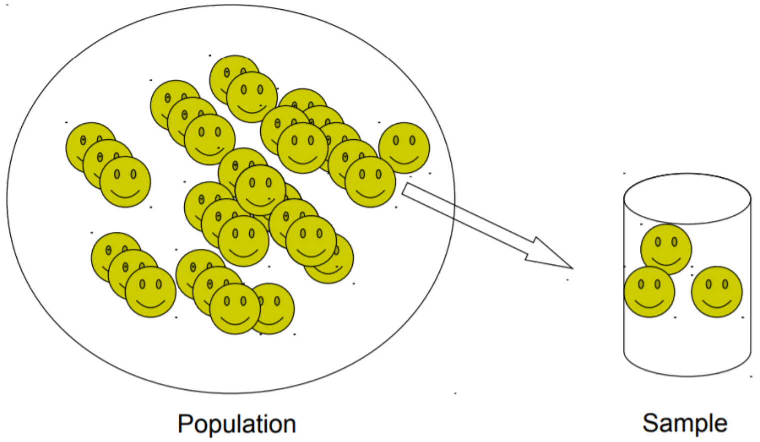
- **Qualitative - M**easurements that each fail into one of several categories. (hair color, ethnic groups and other attributes of the population)
- **Quantitative** - Observations that are measured on a numerical scale (distance traveled to college, number of children in a family, etc.)



- Variables
  - Qualitative
    - Dichotomic
      - Gender, marital status
    - Polynomic
      - Brand of PC, hair color
  - Quantitative
    - Discrete
      - Children in family, Strokes on a golf hole
    - Continuous
      - Amount of income tax paid, weight of a student

Types of Statistics

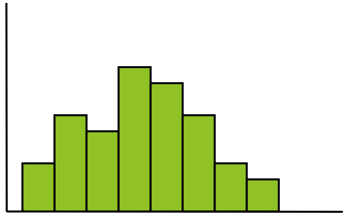
- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest

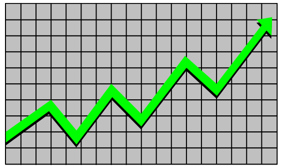


Descriptive Statistics

- Collect data
  - e.g. Survey

- Present data
  - e.g. Tables and graphs





- Summarize data
  - e.g. Sample mean

$$\frac{\sum X_i}{n}$$

Descriptive Statistics

Types of descriptive statistics:

Organize Data

l Tables

l Graphs

Summarize Data

l Central Tendency

l Variation or Dispersion or spread

Summarizing the data

Central Tendency (or Groups’ “Middle Values”)

ü Mean

ü Median

ü Mode

Variation (or Summary of Differences Within Groups)

l Range

l Interquartile Range

l Variance

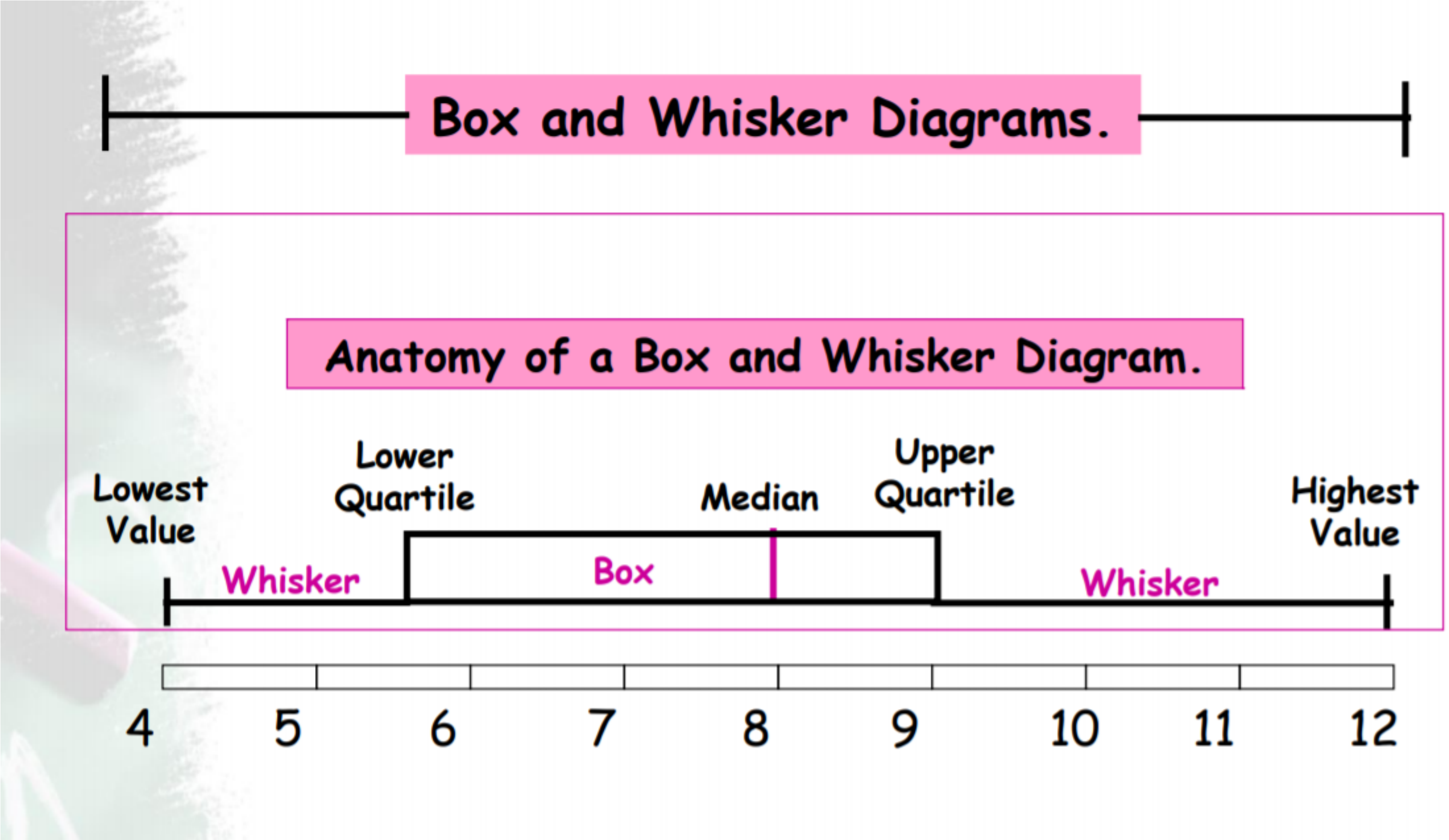
l Standard Deviation

The 5 Number Summary

- The five number summary is another name for the visual representation of the box and whisker plot.

- The five number summary consist of :
  - The median (2nd quartile)
  - The 1st quartile
  - The 3rd quartile
  - The maximum value in a data set
  - The minimum value in a data set

A Box and whisker Plot



Inferential Statistics

- Estimation -
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions on properties of a population based on that of a sample

Sampling

A sample should have the same characteristics as the population it is representing.

Sampling can be:

- **with replacement:** a member of the population may be chosen more than once (picking the candy from the bowl)
- **without replacement:** a member of the population may be chosen only once (lottery ticket)

Sampling methods can be:

- **Random** (each member of the population has an equal chance of being selected**)**
- **Non random**

Random Sampling techniques

- **Simple random sampling -** Each sample of the same size has an equal chance of being selected
- **Stratified sampling** - Divide the population into groups called strata and then take a sample from each stratum
- **Cluster sampling** - Divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.
- **Systematic sampling** - Randomly select a starting point and take every n-th piece of data from a listing of the population
- **Convenience sampling -** Elements in the sample are chosen based on convenience of data collection. This technique faces criticism saying that the sample is often not representative enough.

Intro to Probability

- A statistical experiment
  - outcome happens by chance
  - it can have more than one possible outcome
  - Each possible outcome can be specified in advance by probability value like toss of coin results in head or tail with probability1/2.
- A Sample space is a set of all possible outcomes of statistical Experiment.
- A Sample point is an element in sample space. Each sample point associate sample point probability e.g. toss of coin 1/2 , roll of dice 1/6.
- An event is a subset of sample space having one or more sample points.
- Events can be
  - mutually exclusive if no sample points in common
  - Two events are independent when the occurrence of one does not effect the probability of occurrence of the other.

Intro to Probability

- What is the probability for toss of a coin to be head or tail ?
- When you roll a dice, what is the probability for each sample point ?
- When you roll dice 5 times what is the probability for getting 3 odd numbers ?
- Draw a card and what is the probability of it being spade ?

Probability is likelihood of occurrence of an event. If the event is certain to occur, the probability value is 1, if less chances to occur means its value is near to zero. The value of probability lies in between 0 and 1.

Types of events

- Simple event - An outcome from a sample space with one characteristic
  - eg. A red card from a deck of cards
  - Probability for such an event is called simple(marginal) probability.
- Complement of an event A (denoted A/) - All outcomes that are not part of event A
  - eg. All cards that are not diamonds
- Joint event - Involves two or more characteristics simultaneously
  - eg. An ace that is also red from a deck of cards
  - Probability for such an event is called joint probability.

Random variables

- When a variable X - numerical value is determined by a chance event or statistical experiment , it is called A random variable.
- It's value can be predicted before the experiment like probability of getting head is ½ in tossing a coin.
- Random variables are usually assigned the capital letters X, Y or Z
- Random variables can be either discrete or continuous.

Discrete Random variables

- A discrete random variable is one that can assume only a countable number of values (No fractions) .
- Example –
  - Flipping of a coin and count the no. Of heads. This will be integer value from 0 to Plus infinity without fractions.
  - A multiple choice exam of 20 questions. The random variable X is the number of correct answers.
  - Possible values for X are 0, 1, 2, 3, 4, 5, ……. 20.
- In general, with Discrete Random Variables, we are concerned with counting something.

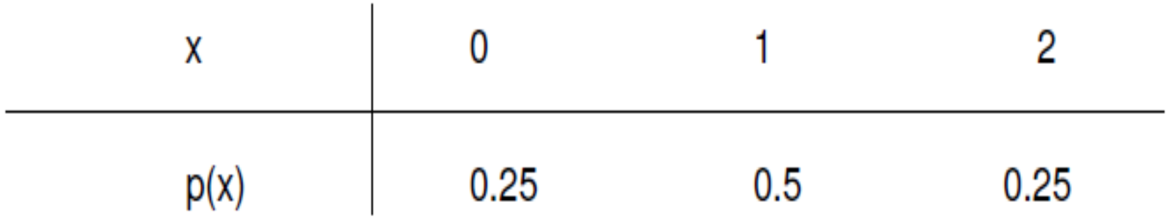
Continuous Random variables

- A Continuous Random Variable can assume any value within a range of values which include fractions like x.99.
- With a Continuous Random Variable, we are generally concerned with measuring something
- Example
  - Flipping coin many times and computing average no. Of heads per flip 100 heads in 150 flips then 0.66 heads per flip
  - The time spent studying for a course per week could be the measurement variable X.
  - It could be measured in days, hours, minutes, seconds, etc. (say 600 minutes/week, or 591 minutes/week, or 590 minutes and 45 seconds, and so on)

Discrete probability distribution

- A Discrete Probability Distribution describes how the probabilities are distributed over the various values that the discrete random variable can take
- The probability distribution for the discrete random variable X, is a table, graph or formula that gives the probability of observing each value of x
- We denote the probability of each x by the symbol p(x)
- 2 important rules for probability distributions;
  - $0 < p(x) < 1$ for all values of x
  - $\Sigma p(x) = 1$ (sum of probabilities of all sample points is equal to 1.)

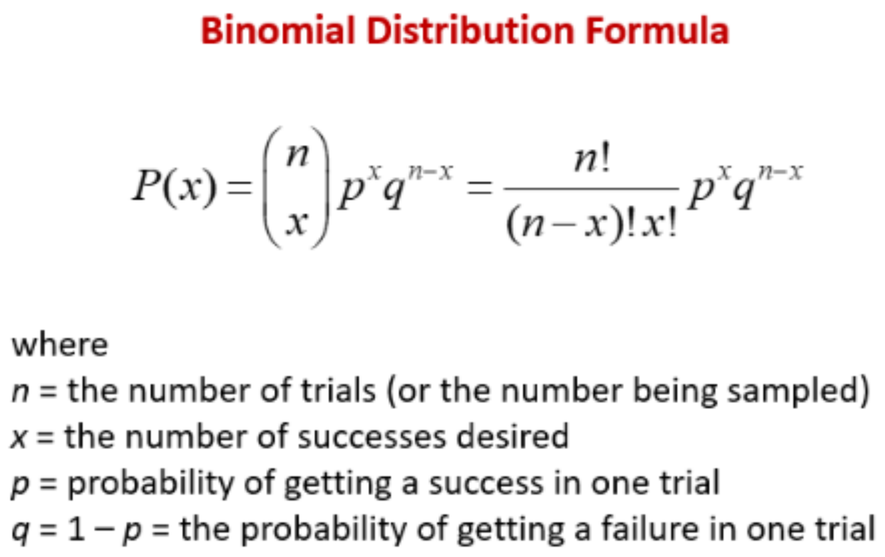
Example: Toss two coins…….Let X be defined as the number of heads occurring in the two tosses.



x	0	1	2
p(x)	0.25	0.5	0.25

The Binomial Distribution

- The Binomial Distribution is used to describe the response from a Binomial Experiment
- In a Binomial Experiment there are only two possible outcomes.
  - Example: Yes/no, for/against, present/absent, +/-, will buy/will not etc.
  - These outcomes are usually referred to as success/ failure
- A Binomial Experiment usually consists of a number of repeated trials



Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

TRY THIS

There are 10 questions with multiple choice answers a, b, c, d, e. Find out the probability of 4 correct answers.

Test your understanding

Let's say that a dice is rolled 5 times. Let's also define a random variable X which is supposed to count the number of dice rolls that produced odd numbers.

1. What is the probability that more than half the 5 rolls of the dice will be odd?
2. What is the probability that there will be more than 1 and less than 4 odd numbers?

The trick here is to write out the probability distribution correctly

Probability Density Function

Probability Density function (PDF) gives the probability values that occur at a given point in the outcome space. In general it refers to continuous random variable.

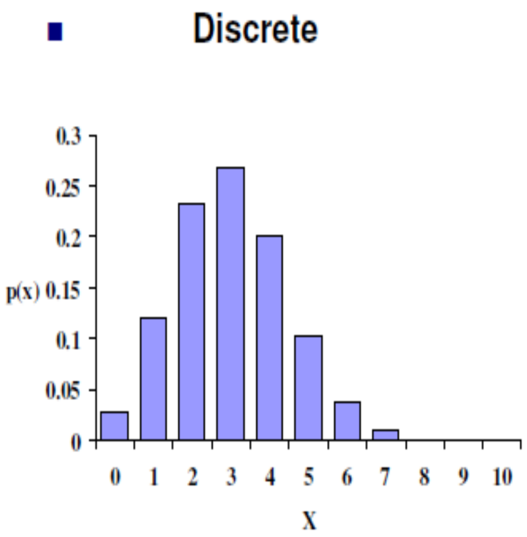
Probability distribution function could be either cumulative distribution or probability mass function. In rare occasions, it denotes PDF. This refers to a range of values that probability could take in the interval.

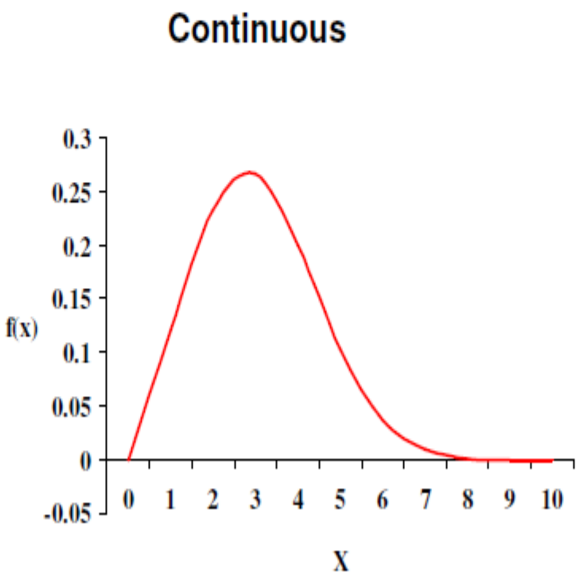
Cumulative distribution gives the value for above or below a limit ex. Ogive.

Probability Mass function gives the probability that a discrete random variable is exactly equal to some value

Continuous Probability distribution

- In discrete random variable distributions, there is always a gap between the points of a distribution. i.e.. the number of odd dice in five rolls can only be 0, 1, 2, 3, 4 or 5, not 4.2 or 3.5, etc.
- In a continuous distribution, there are no gaps between values
- The probability density function f(x) must satisfy two conditions;
  - f(x) >= 0, (i.e. non negative)
  - The total area under the curve is 1 (compare this with P(x) = 1 for discrete)

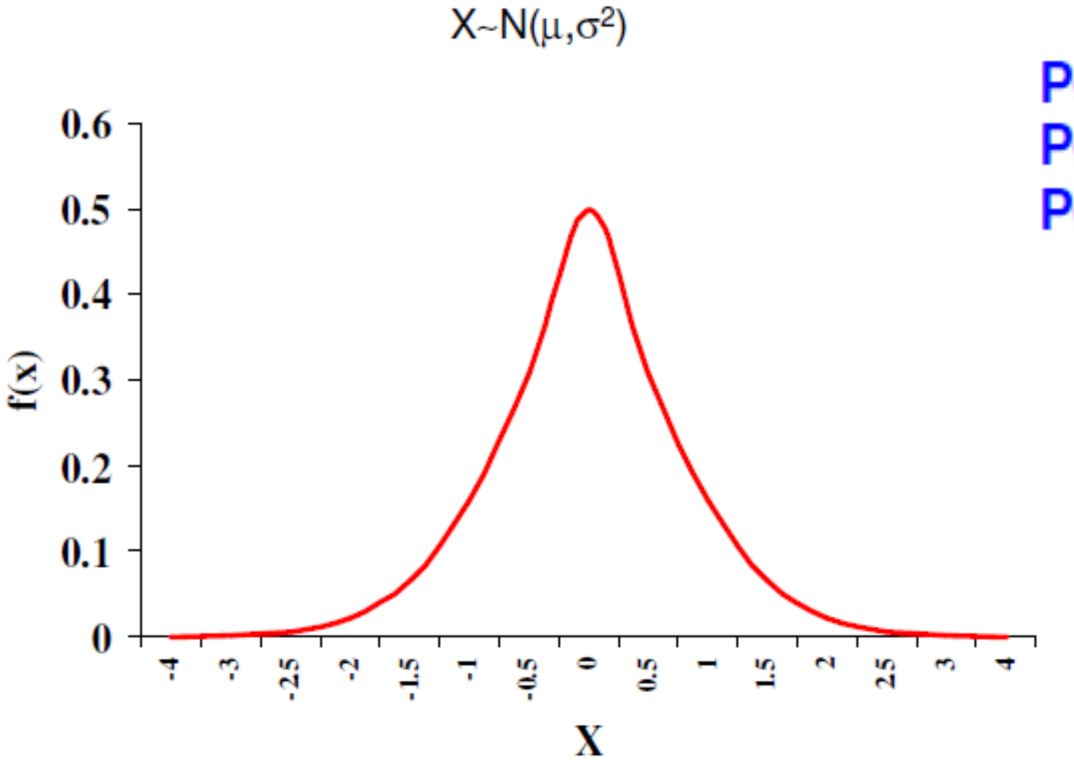




Normal Distribution

- The most important continuous probability distribution is the Normal Distribution.
- The reason is that it has a very important use in the statistical theory of drawing conclusions from sample data about the populations from which the samples are drawn, and in Statistical Process Control.
- It can be determined entirely by the values of μ and Sigma.
- There are several characteristics that make the normal distribution very important for statisticians:
    - It is bell shaped
    - Symmetrical about Mean which is also Median and Mode
    - Most observations in the distribution are close to the mean, with gradually fewer observations further away

Normal Distribution



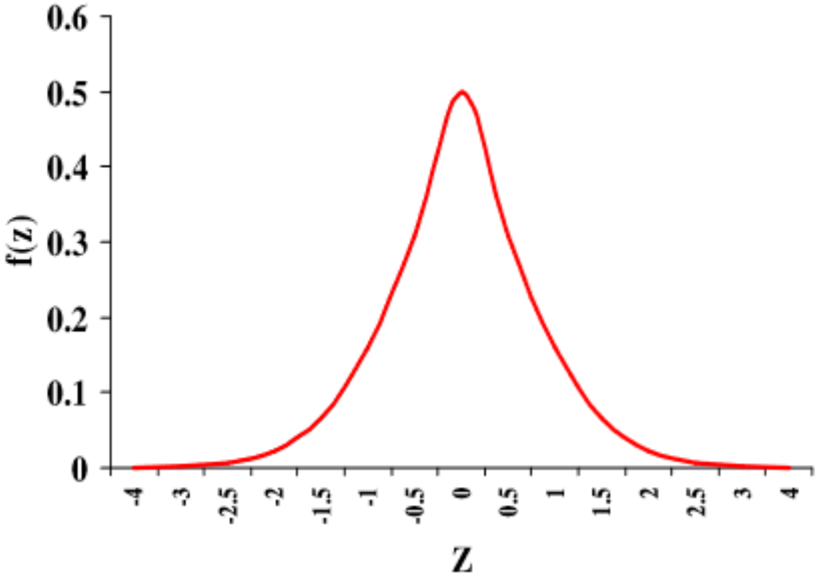
$P(\mu-\sigma < X < \mu+\sigma) = 0.683$

$P(\mu-2\sigma < X < \mu+2\sigma) = 0.954$

$P(\mu-3\sigma < X < \mu+3\sigma) = 0.997$

Standard Normal Distribution

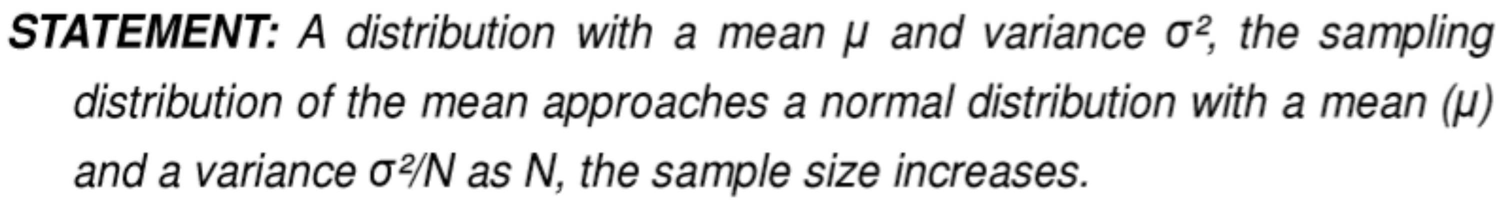
- A special case of the normal distribution, the standard normal distribution has a mean of 0 and a standard deviation of 1
- The corresponding standard random variable is denoted by Z
- One of the most popular applications – Outlier detection via z-score method



Any normal distribution can be converted to the Standard Normal Distribution, simply by converting it’s mean to 0 and it’s standard deviation to 1

$$Z = \frac{X - \mu}{\sigma}$$

Central limit theorem



STATEMENT: *A distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size increases.*

- The amazing and counter-intuitive thing about the central limit theorem is that the distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal
- If enough samples are taken repeatedly from a population, the Centre of the distribution of the sample means, is μ, the population mean
- If the underlying population has a large variance, then naturally the sample means will also have a large variance
- As the sample size n increases, the variance of the sampling distribution decreases. This is logical, because the larger the sample size, the closer we are to measuring the true population parameters

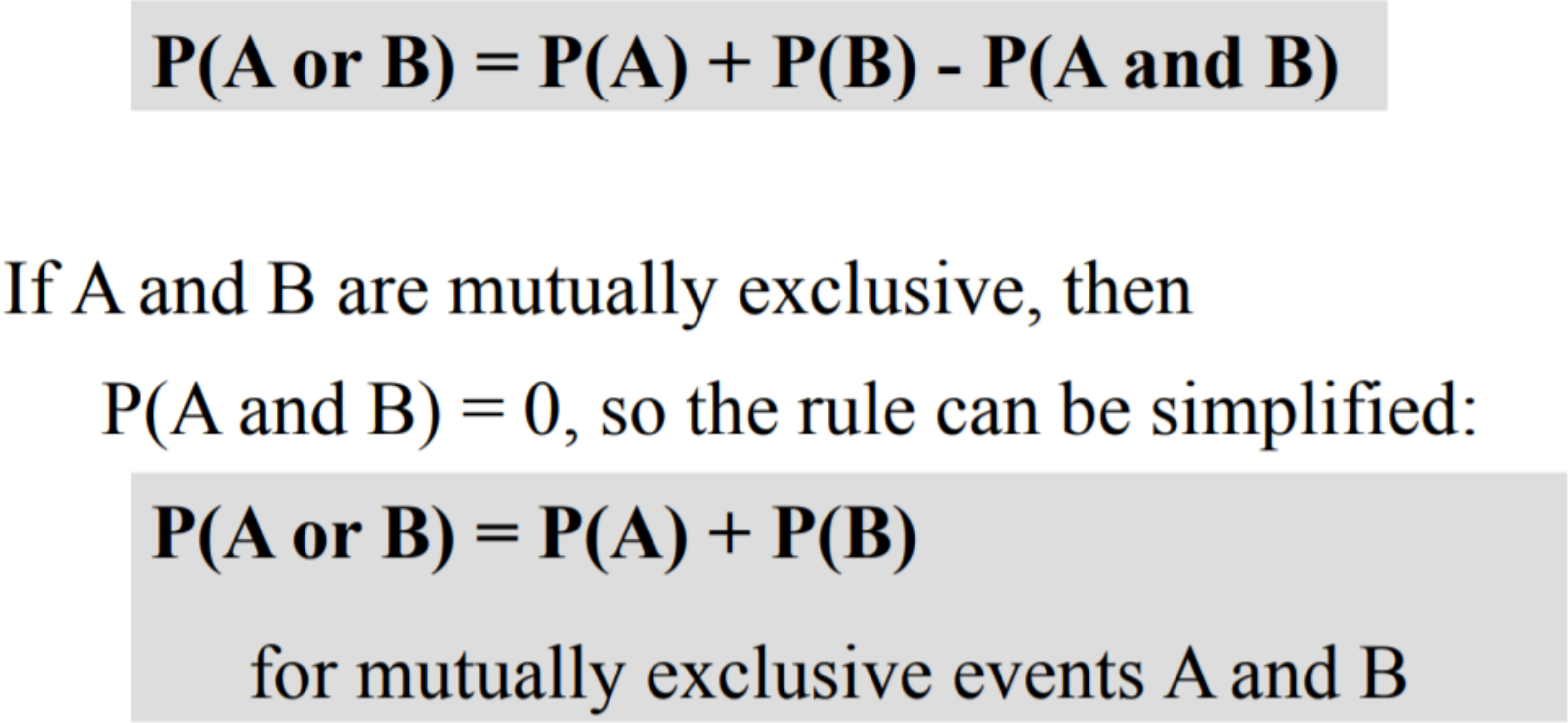
Assumptions - CLT

- The **data must follow the randomization condition**. It must be sampled randomly
- **Samples should be independent of each other.** One sample should not influence the other samples
- The **sample size should be sufficiently large**. Now, how we will figure out how large this size should be?

In general, **a sample size of 30 is considered sufficient when the population is symmetric**.

Some Probability Principles

General Addition rule -



$$\mathbf{P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)}$$

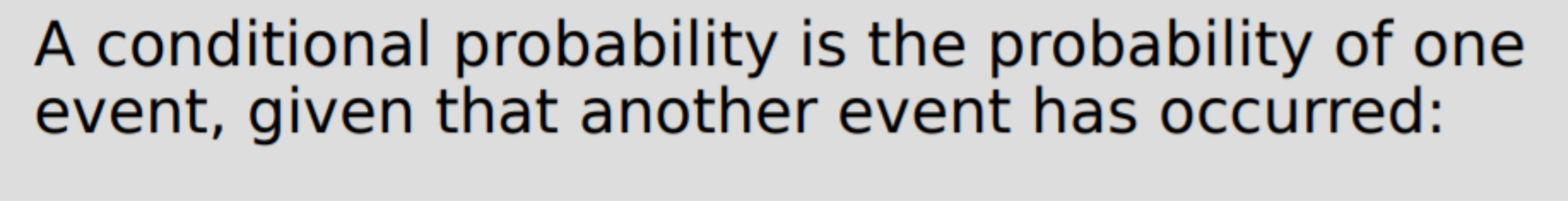
If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified:

$$\mathbf{P(A\ or\ B) = P(A) + P(B)}$$

for mutually exclusive events A and B

Some Probability Principles



A conditional probability is the probability of one event, given that another event has occurred:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ and } B) = P(A \mid B)\,P(B)$$

This is often referred to as the multiplication rule

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

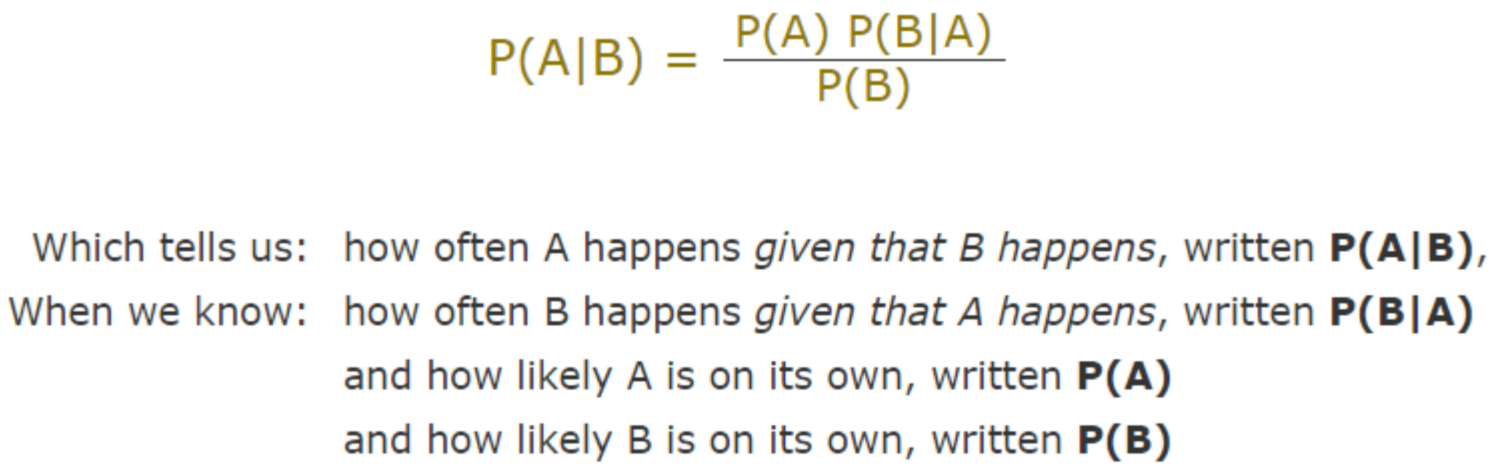
Test Your Understanding

Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC ?

Suppose a city council is composed of 10 democrats, 7 republicans, and 2 independents. Find the probability of randomly selecting a democrat followed by an independent.

Bayes Theorem

Developed by Thomas Bayes in the 18th Century. It is an extension of conditional probability.



$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

Which tells us:	how often A happens *given that B happens*, written **P(A|B)**,
When we know:	how often B happens *given that A happens*, written **P(B|A)**
	and how likely A is on its own, written **P(A)**
	and how likely B is on its own, written **P(B)**

Test Your Understanding

You are planning a picnic today, but the morning is cloudy

1. Oh no! 50% of all rainy days start off cloudy!
2. But cloudy mornings are common (about 40% of days start cloudy)
3. And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

The Art Competition has entries from three painters: Pam, Pia and Pablo

1. Pam put in 15 paintings, 4% of her works have won First Prize.
2. Pia put in 5 paintings, 6% of her works have won First Prize.
3. Pablo put in 10 paintings, 3% of his works have won First Prize.

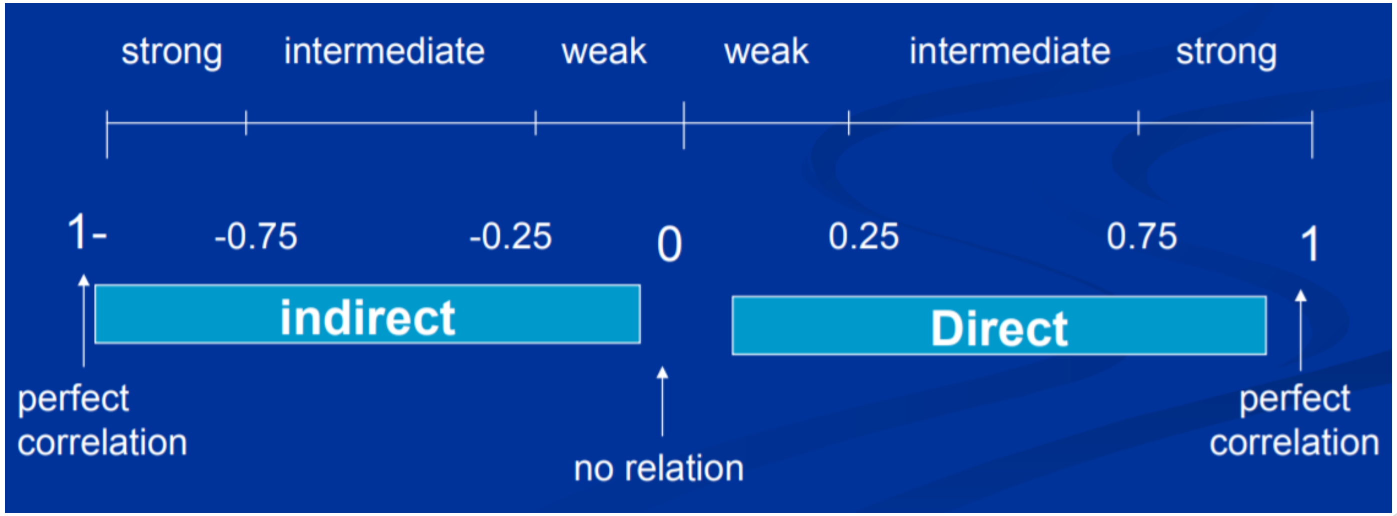
What is the chance that Pam will win First Prize?

Correlation

- The strength of the relationship between two variables is measured by the coefficient of correlation coefficient r, ‘rho’.
- Correlation coefficients range between -1 and +1 and zero means that relationship is not linear but cannot say that there is no relationship . They may have strong curvilinear relationship.
- -ve correlation coefficients indicate negative relationships. i.e. as one variable increases, the other decreases
- Stronger linear relationships have values closer to ± 1, weaker linear relationships have values closer to 0.
- 0 indicates no relationship at all and the relationship is not linear.
- ± 1 indicates a perfect relationship
  - E.g.: Income and expenditure is positive
  - E.g.: Price and demand of commodity is negative

Mathematical correlation

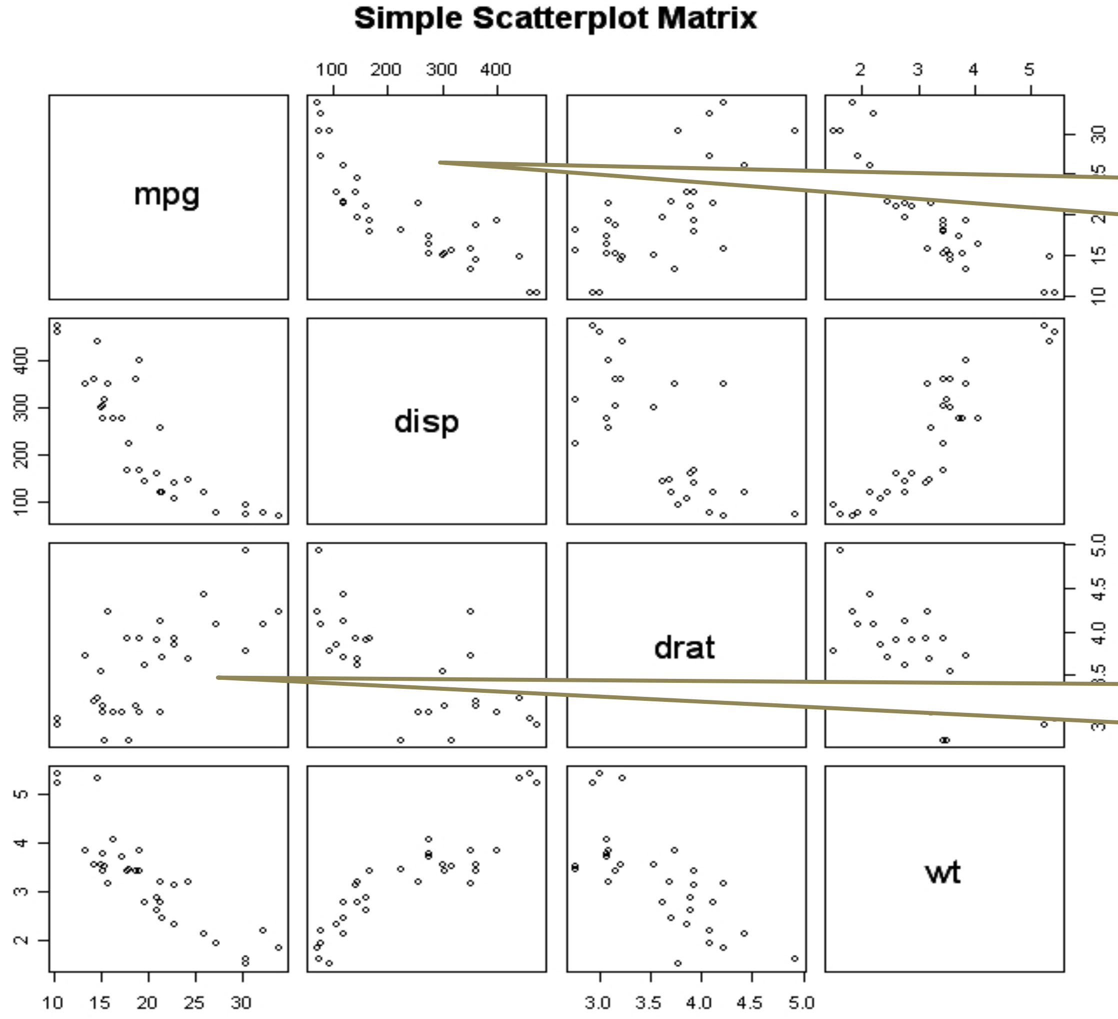
$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^{2} - \frac{(\sum x)^{2}}{n}\right)\cdot\left(\sum y^{2} - \frac{(\sum y)^{2}}{n}\right)}}$$



Correlation and Causation

- Correlation analysis helps determine degree of relationship between two or more variables
- It does not tell about cause and effect relationship
- Even high degree of correlation does not necessarily mean a relationship of cause and effect exists between variables
- Correlation does not imply causation though the existence of causation always imply correlation
- Examples
  - More firemen are there so the fire is big but the fire is not caused by Firemen.
  - When one sleeps with shoes on, he is likely to get headache. This may be due to alcohol intoxication.

Correlation – Using Scatter plots



Looks like a strong case for negative interdependence

This is a good example for positive correlation

Estimation in Inferential Statistics

- If we take a sample from a population, we can estimate parameters from the population, using sample statistics
- Example:
  - The sample mean (x) is our best estimate of the population mean (μ)
- If we estimate a range or interval within which the true population parameter lies, then we are using an interval estimation method
- This is the most common method of estimation. We can also apply a level of how confident we are in the estimate
- Point estimate
  - The mean annual rainfall of Melbourne is 620mm per year
- Interval Estimate
  - In 80% of all years Melbourne receives between 440 and 800 mm rain

Hypothesis Testing

- A statistical Hypothesis is an assertion regarding the statistical distribution of the population and validating this assertion.
- It is a statement regarding the parameters of the population.
- Testing of Hypothesis deals with the verification of validity of presumption regarding the parameters of the population using samples drawn from the population.
- Statistical Hypothesis is denoted by H.
- In a Test Procedure, to start with, a hypothesis is made. The validity of the hypothesis is tested.
- If the hypothesis is found to be true, it is accepted. If it is found to be untrue, it is rejected.
- The hypothesis which is being tested for possible rejection is called null hypothesis
- Null hypothesis is denoted by H0 The hypothesis which is accepted when null hypothesis is rejected is called alternate Hypothesis Ha

Hypothesis Testing - Steps

Lets look at the key steps that are involved in this process

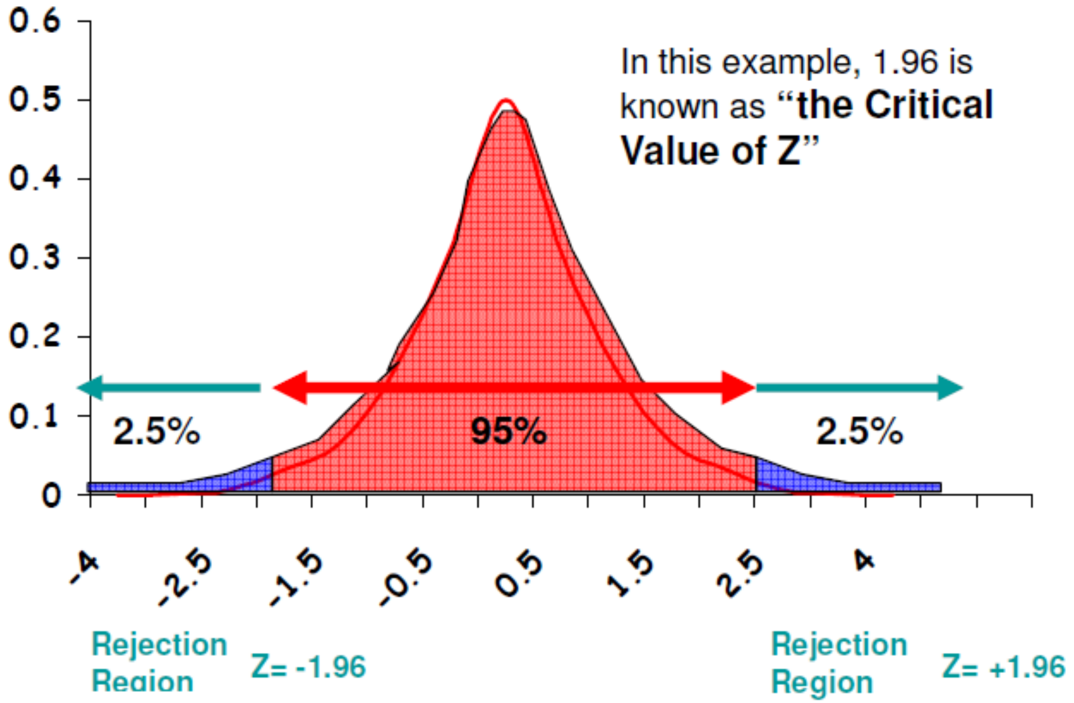
- Null hypothesis
- Alternative hypothesis
- Confidence level
- Decision Rule
- Test statistic
- Decision

In any hypothesis test, we must specify a null or ‘no effect’ hypothesis before we perform the test. We always assume the null hypothesis is true, or at least is the most plausible explanation before we do the test. The test can only disprove the null hypothesis.

Hypothesis Testing - Decision rule

After we know the H_0 null and H_1 alternative hypotheses and the level of confidence α associated with the test, we determine the points on the distribution of the test statistic where we will decide when the null hypothesis should be rejected in favor of the alternative hypothesis

Using P value/Critical Value or Z alpha for single side , Z alpha/2 for two side.



In this example, 1.96 is known as **"the Critical Value of Z"**

Type-I and Type-II error

Process of testing a hypothesis indicates that there is a possibility of making an error. There are two types of errors:

- Type I error: The error of rejecting the null hypothesis H0 even though H0 was true.
- Type II error: The error of accepting the null hypothesis H0 even though H0 was false.

For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

- P (type I error) = α
- P (type II error) = $1 - \alpha$

Level of significance and Rejection region

The type I error just is the significance level of the test. In other words, The significance level is the fixed probability that the null hypothesis will be rejected when it is true

- Significance Level = P (type-I error) = α
- In statistics, a result is called significant if it is unlikely to have occurred by chance.
- Usually, the significance level is chosen to be 0.05 (or equivalently, 5%).

The rejection region is the area (s) determined by the decision rule

- Say, we decide to perform a z test with 95% confidence (a = level of significance = 0.05)
- We know that 95% of z values lie within 1.96 standard deviations of the mean. Therefore we may decide that observations outside this range are “statistically significant” at the 0.05 level
- In other words, we will reject the null hypothesis if the test statistic z is > +1.96 or <-1.96

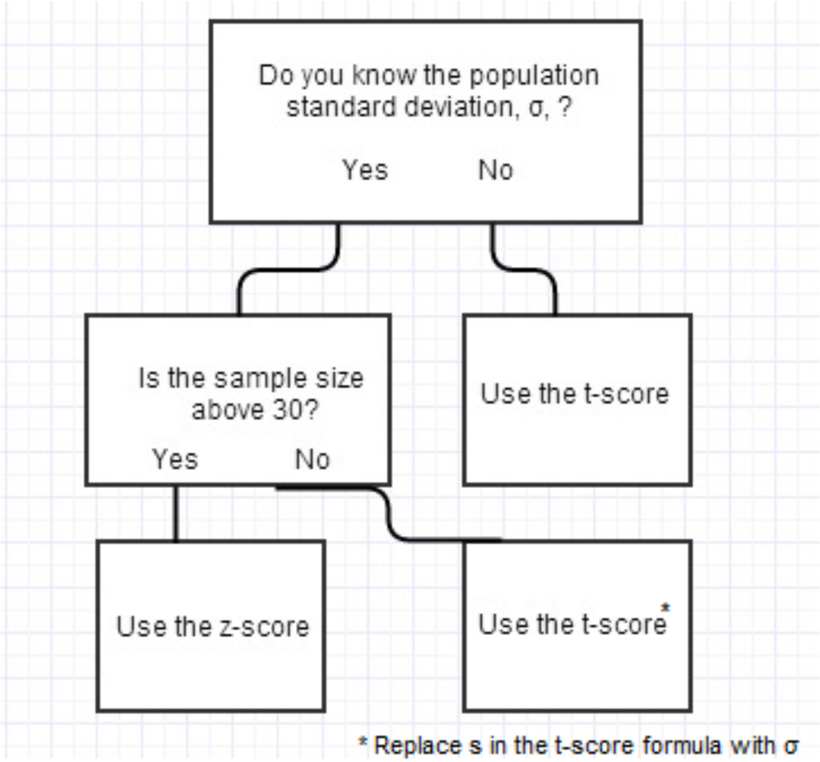
Hypothesis Tests - Example

Problem : Suppose that we have been told that the price of petrol in Melbourne is normally distributed with a mean of 92 cents per litre, and a standard deviation of 3.1 cents/litre. To test whether this price is in fact true, we sample 50 service stations and obtain a mean of 93.6 cents/litre

Problem : A company pays production workers $630 per week. The union claims that these workers are paid below the industry average for their work. A sample of 15 workers from other sites gives a mean wage of $670/week with a standard deviation of $58/week. Is the unions claim justified?

T test and Z test

T-test refers to a univariate hypothesis test based on t-statistic, wherein the mean is known, and population variance is approximated from the sample. On the other hand, **Z-test** is also a univariate test that is based on standard normal distribution.



Do you know the population standard deviation, σ, ?

Yes No

Is the sample size above 30?

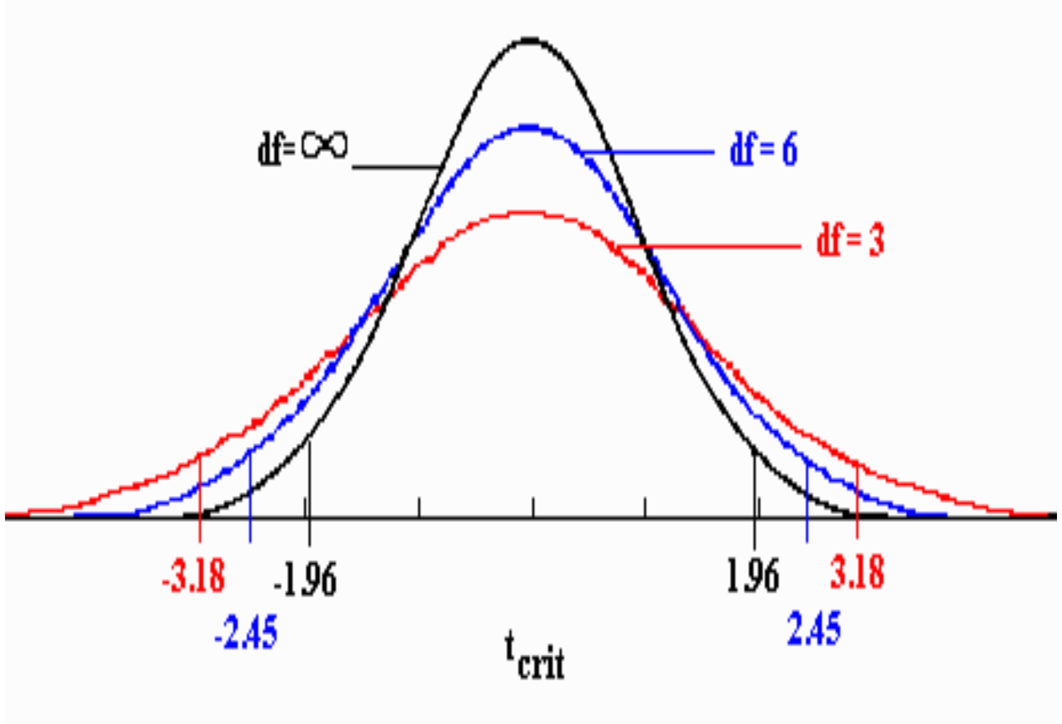
Yes No

Use the t-score

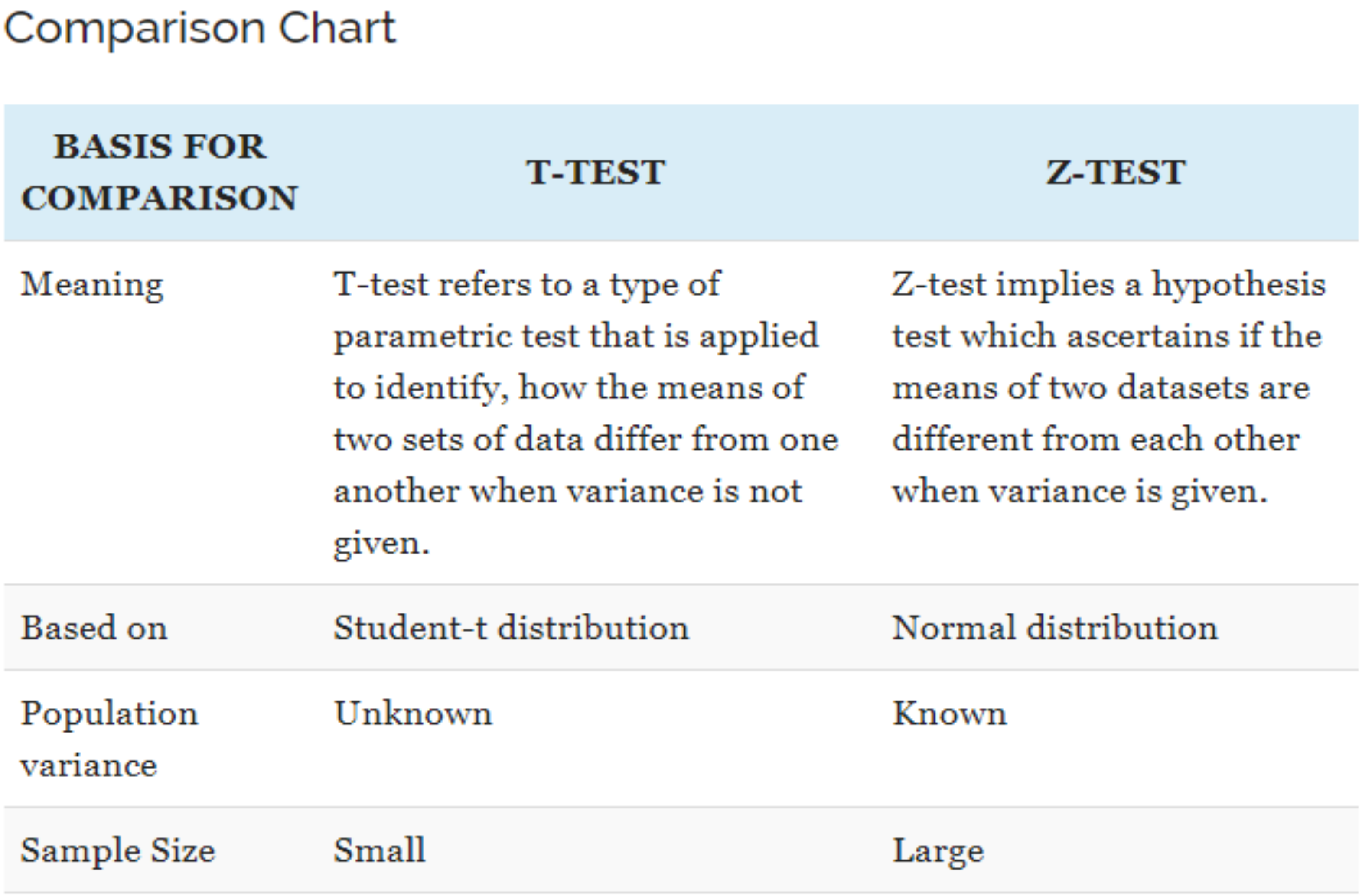
Use the z-score

Use the t-score*

* Replace s in the t-score formula with σ



T test and Z test



Comparison Chart

BASIS FOR COMPARISON	T-TEST	Z-TEST
Meaning	T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given.	Z-test implies a hypothesis test which ascertains if the means of two datasets are different from each other when variance is given.
Based on	Student-t distribution	Normal distribution
Population variance	Unknown	Known
Sample Size	Small	Large

Linear Regression

Regression (Y = mX + C where C is intercept , m is slope of X)

- ❑ Quantifying the relationship between two continuous variables
- ❑ Predict (or forecast) the value of one variable from knowledge of the value of another variable
- ❑ That is an estimating equation - a mathematical formula - will be developed

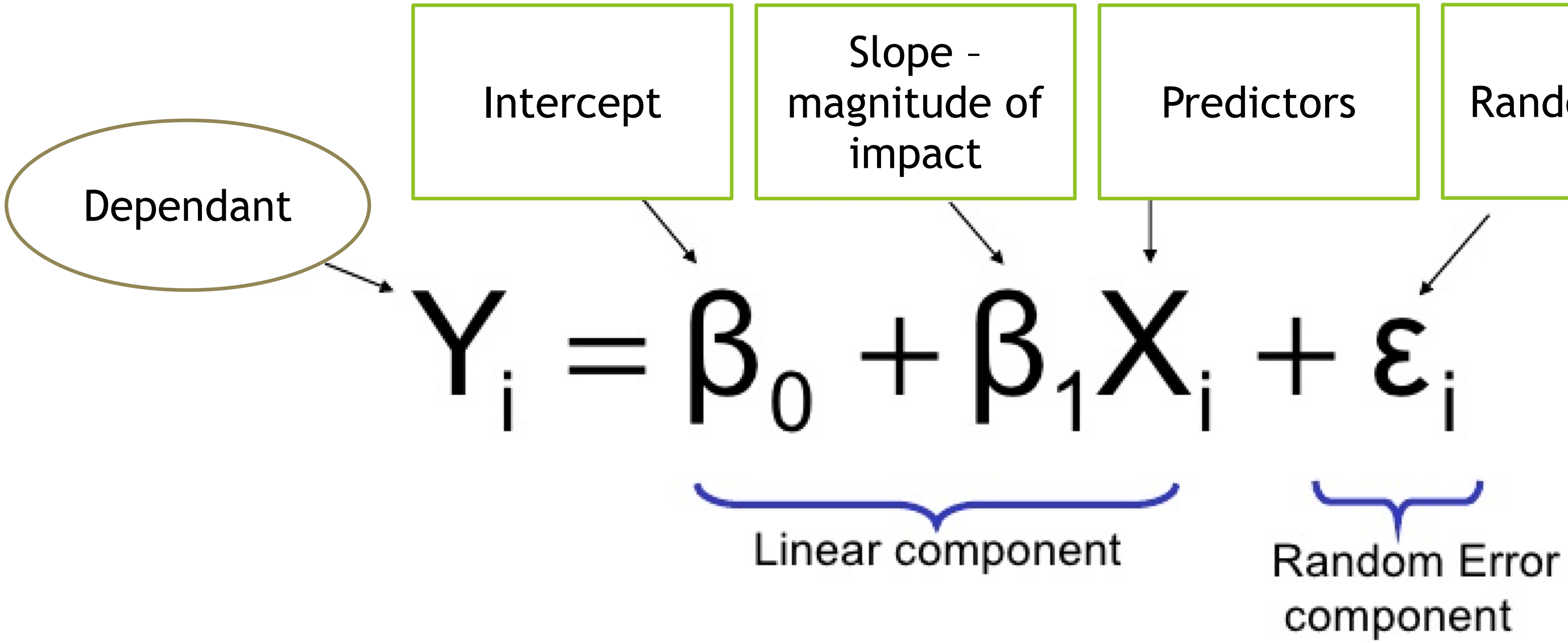
Correlation (coefficient r lies between -1-0-+1, 0 means relations is not linear)

- ❑ When the pattern of relationship is known, Correlation analysis can be applied to determine the degree to which the variables are related
- ❑ Correlation analysis informs how well the estimating equation actually describes the relationship
- ❑ For Stronger correlation r should be greater or less than 0.5

Nature of Dependence

- Deterministic Relationship
- Lease of SUV over the weekend:
  - Fixed Cost $250.00
  - Plus $0.40/mile.
  - X = # of miles you drive SUV
  - Y = total lease cost
  - $Y = 250 + 0.40\ X$
  - No work for a statistician to do here—there's nothing random (stochastic) about them.
- Stochastic Relationship
  - Where there is a random element—where you can't predict with absolute certainty Y for a given X.

Linear Regression



Dependant

Intercept

Slope – magnitude of impact

Predictors

Random error

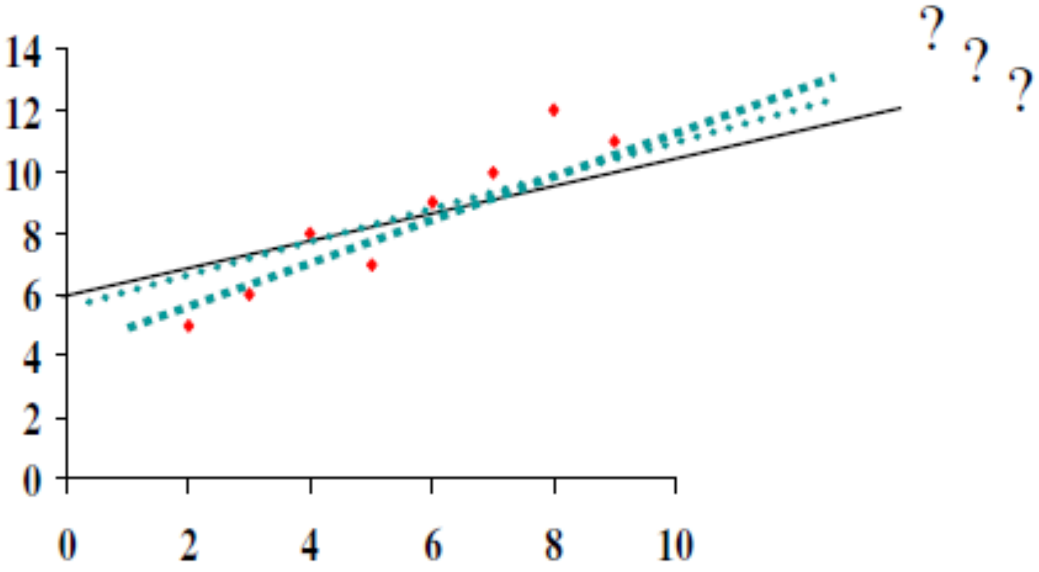
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

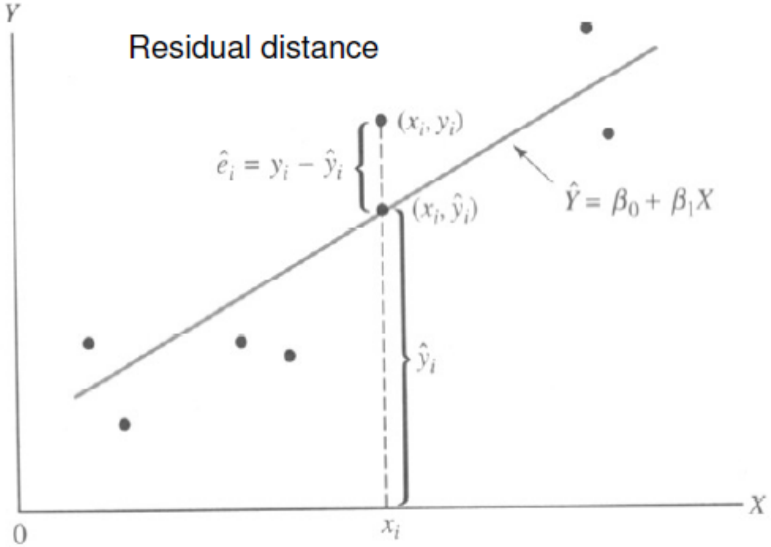
Linear Regression : Continued

If we were to express what regression does in one line

“Regression helps us get to the line that best fits a linear data distribution”

Because the line will seldom fit the data precisely, there is always some error associated with our line. The line of best fit is the line that minimizes the spread of these errors

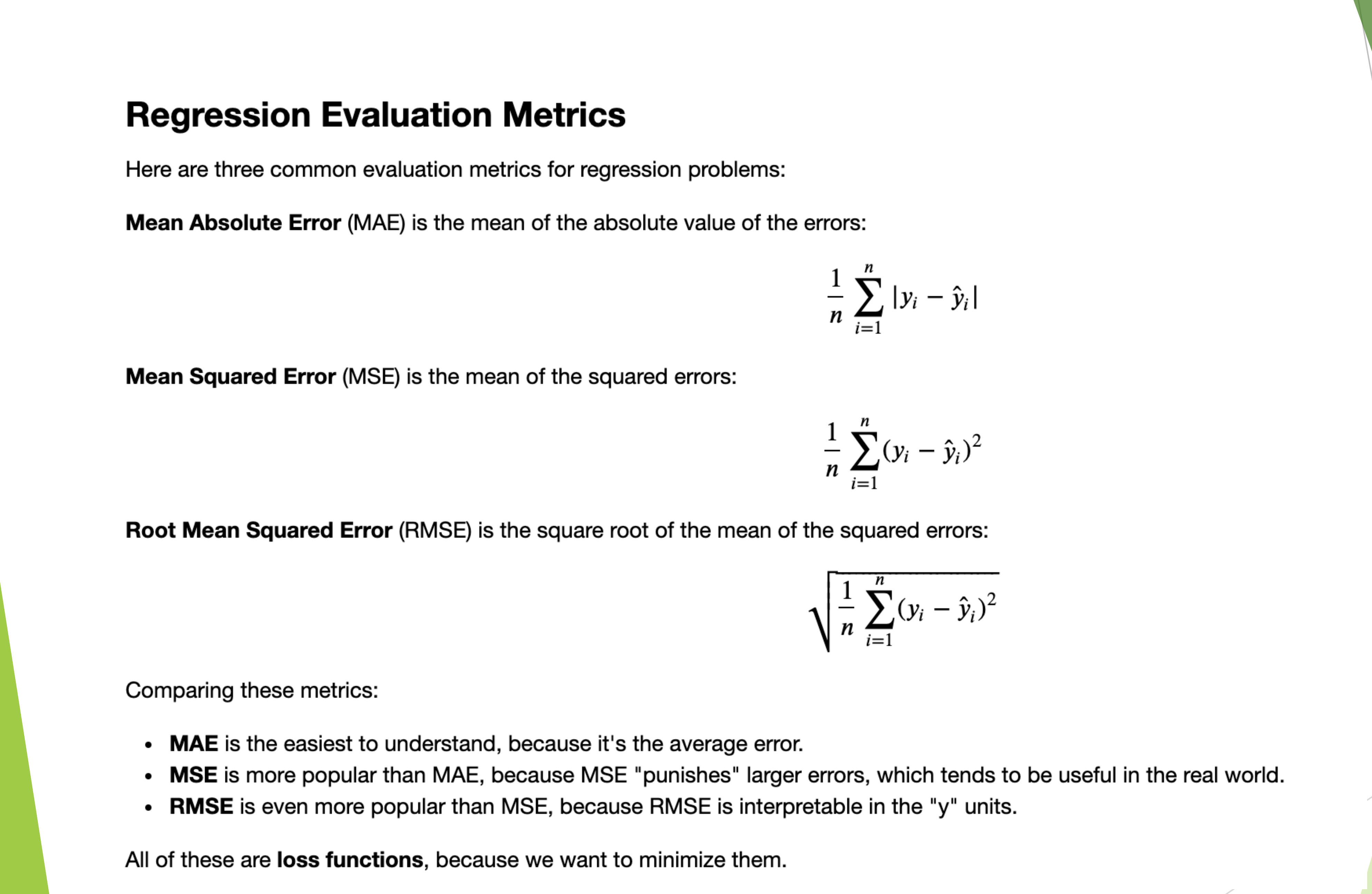




Linear Regression : Assumptions of OLS

Lets look at the base assumptions for the error term that are mandatory in a Linear regression when solved using the Ordinary least squares technique

- The error variable is normally distributed - Required for hypothesis test not OLS estimates
- The expected value of the error variable is zero
- The variance of the error is constant over the entire range of X values – Homoscedasticity
- The errors associated with any two Y values are independent - Autocorrelation
- Coefficients are linear
- Multicollinearity must be avoided at all costs



Regression Evaluation Metrics

Here are three common evaluation metrics for regression problems:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Comparing these metrics:

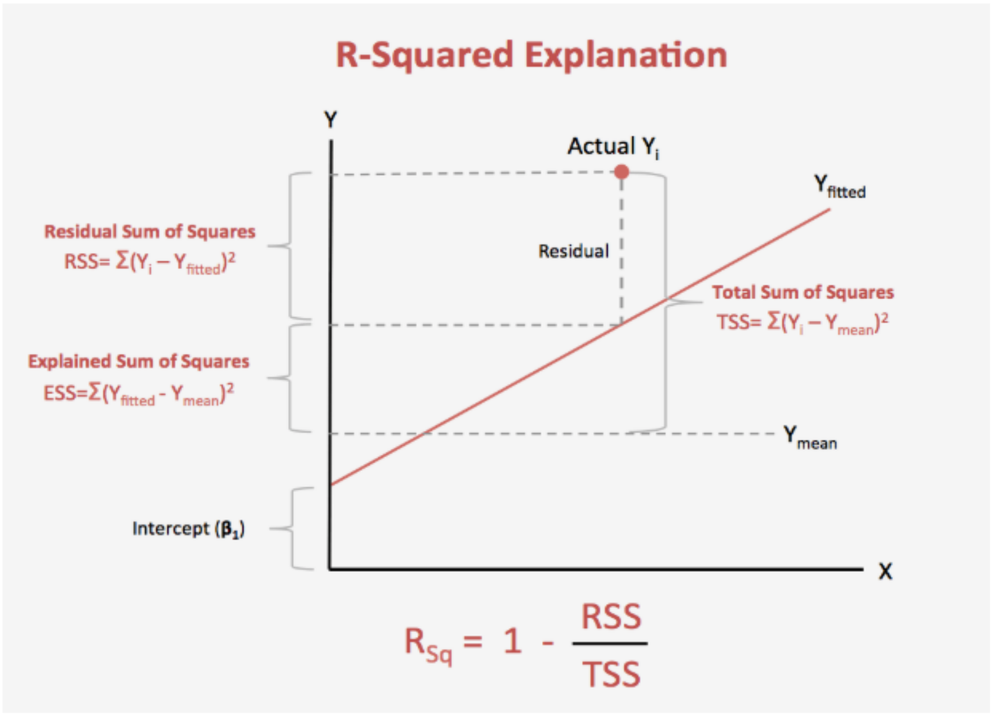
- **MAE** is the easiest to understand, because it's the average error.
- **MSE** is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- **RMSE** is even more popular than MSE, because RMSE is interpretable in the "y" units.

All of these are **loss functions**, because we want to minimize them.

Linear Regression - Goodness of Fit

R-Squared and Adjusted R-Squared

What R-Squared tells us is the proportion of variation in the dependent (response) variable that has been explained by this model.



R SQUARED COMPUTATION

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i}^{n} (y_i - \hat{y_i})^2$$

$$TSS = \sum_{i}^{n} (y_i - \bar{y_i})^2$$

Linear Regression : Goodness of Fit

What is adjusted R-Squared?

As you add more X variables to your model, the R-Squared value of the new bigger model will always be greater than that of the smaller subset.

Why is that ??

This is because, since all the variables in the original model is also present, their contribution to explain the dependent variable will be present in the super-set as well.

It is here, the adjusted R-Squared value comes to help.

- ❑ Adjusted R-Squared is formulated such that it penalises the number of terms (read predictors) in your model.
- ❑ So unlike R-squared, as the number of predictors in the model increases, the adjusted R-squared may not always increase.

Logistic Regression - History

- Until 1972, people didn't know how to analyze data which had a non-normal error distribution in the dependent variable
- In 1972, came a breakthrough by John Nelder and Robert Wedderburn in the form of **Generalized Linear Models**.

Key tenets of GLMs

- These models comprise a linear combination of input features.
- The mean of the response variable is related to the linear combination of input features via a link function.
- The response variable is considered to have an underlying probability distribution belonging to the family of exponential distributions such as binomial distribution, Poisson distribution, or Gaussian distribution.

Logistic Regression

In linear regression the Y variable is always a continuous variable. If suppose, the Y variable was categorical, you cannot use linear regression model it. So what would you do when the Y is a categorical variable with 2 classes?

Logistic regression can be used to model and solve such problems, also called as binary classification problems.

Important Points :

- Dependent variable is a categorical dichotomy :*Y can have 2 classes only and not more than that. If Y has more than 2 classes, it would become a multi class classification and you can no longer use the vanilla logistic regression for that.*
- Explanatory variables(x) can be either continuous or categorical

Logistic Regression - Key ideas

As discussed Logistic regression is a special case of GLMs

- The response variable must follow a binomial distribution.
- Logistic Regression assumes a linear relationship between the independent variables and the link function (logit).
- The dependent variable should have mutually exclusive and exhaustive categories.

Do we need Logistic regression ??

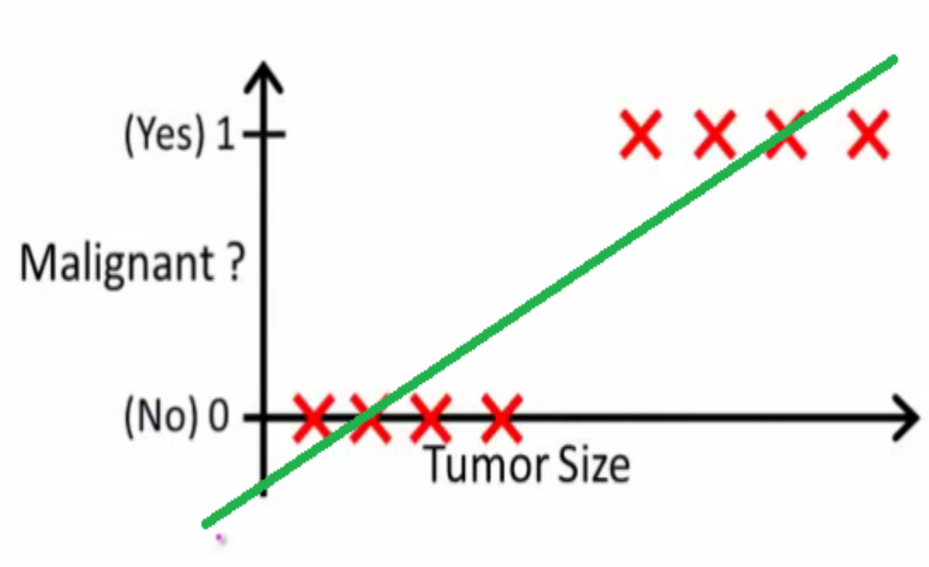


Figure 1

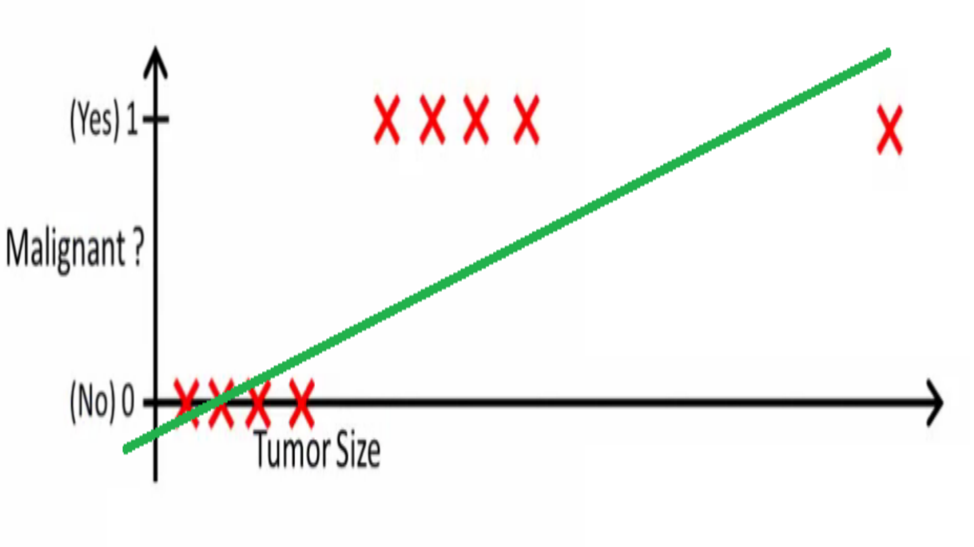


Figure 2

Why cant we use some decision rule and stick with linear regression ??

Logistic Regression

Lets look at a practical example..

We are provided a sample of 1000 customers. We need to predict the probability whether a customer will buy a given magazine or not. Lets say it depends on the age of the customer.

$$g(y) = \beta o + \beta(Age)$$

g() is the link function . This is a standard approach when writing equations for a GLM. This function is established using two things : Probability of Success(p) and Probability of Failure(1-p). p should meet following criteria:

- It must always be positive (since p >= 0)
- It must always be less than equals to 1 (since p <= 1)

Logistic Regression

We need to ensure that the probability is always positive so lets use the exponential function.

p = exp(ßo + ß(Age)) = e^(ßo + ß(Age))

We now need to ensure that these prob values are less than 1

p = exp(ßo + ß(Age)) / exp(ßo + ß(Age)) + 1

= e^(ßo + ß(Age)) / e^(ßo + ß(Age)) + 1

(1 – p) can be calculated similarly and we can then compute p/(1-p)

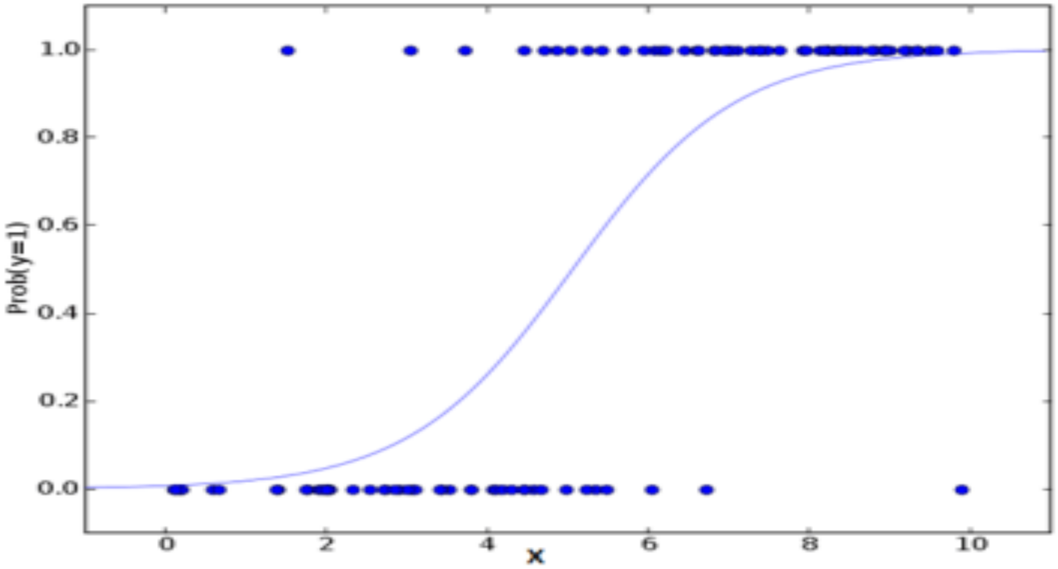
$$\frac{p}{1 - p} = e^{y}$$

Logistic Regression

Taking log on both sides of the equation gives us

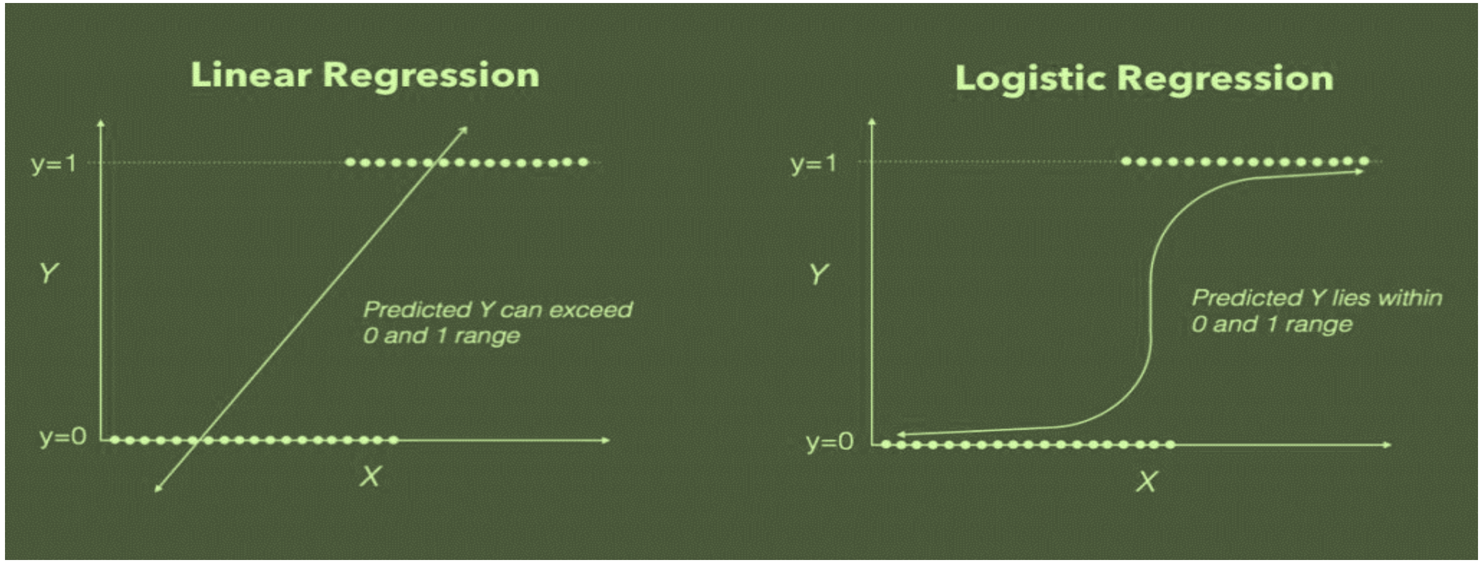
$$\log\left(\frac{p}{1 - p}\right) = \beta_o + \beta(\text{Age})$$

log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.



Why not linear regression ?

- When the response variable has only 2 possible values, it is desirable to have a model that predicts the value either as 0 or 1 or as a probability score that ranges between 0 and 1.
- Linear regression does *not* have this capability. Because, If you use linear regression to model a binary response variable, the resulting model may not restrict the predicted Y values within 0 and 1.



- In logistic regression, you get a probability score that reflects the probability of the occurrence of the event.

Logistic Regression : Estimation

- In Multiple Regression, we use the **Ordinary Least Square (OLS)** method to determine the best coefficients to attain good model fit.
- In Logistic Regression, we use **Maximum Likelihood method** to determine the best coefficients and eventually a good model fit.

How does MLE work ??

- It tries to find the value of coefficients (ßo,ß1) such that the predicted probabilities are as close to the observed probabilities as possible.
- In a binary classification case , maximum likelihood will try to find values of ßo and ß1 such that the resultant probabilities are closest to either 1 or 0. The likelihood function is written as

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Logistic Regression : Evaluation

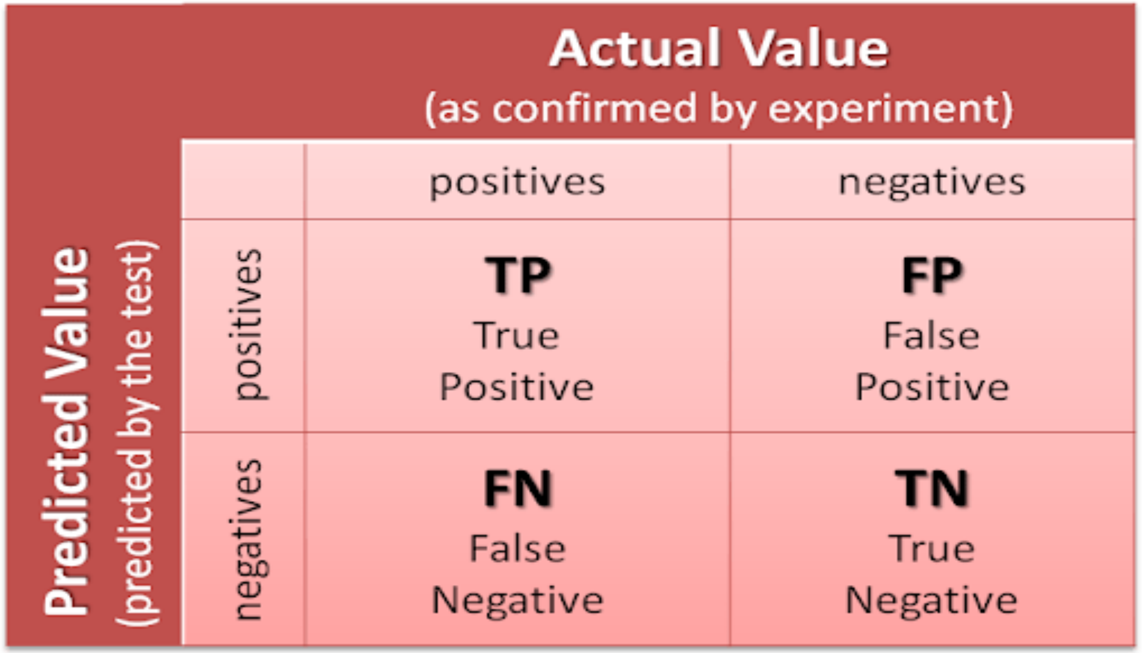
In Linear Regression, we check adjusted R^2, F Statistics, RSE, and RMSE to evaluate model fit and accuracy.

Logistic regression has different metrics to evaluate a model

- **Akaike Information Criteria (AIC)**
  - AIC is just one of several reasonable ways to capture the trade-off between goodness of fit (which is improved by adding model complexity in the form of extra explanatory variables) and parsimony (Simpler is Better) in comparing models
  - This is a direct replacement of the Adjusted R squared value that we use to compare linear models

Logistic Regression : Evaluation

- Confusion Matrix



		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

Accuracy = (TP + TN)/(TP + FP + FN + TN)

True negative rate / Specificity = (TN)/(TN + FP)

True positive rate / Sensitivity = (TP)/(TP + FN)

Logistic Regression : Use cases

Examples of binary classification problems:

- **Spam Detection** : Predicting if an email is Spam or not.
- **Credit Card Fraud** : Predicting if a given credit card transaction is fraud or not
- **Health** : Predicting if a given mass of tissue is benign or malignant
- **Marketing** : Predicting if a given user will buy an insurance product or not
- **Banking** : Predicting if a customer will default on a loan.

Clustering

Clustering is a technique for finding similar groups in data, called clusters.

It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

How do we define “Similar” in clustering?

Measures of Distance

There are three popular measures of Distance

- Euclidean distance - Distance computed on the basis of the Pythagorean distance formula developed in the cartesian coordinate system. This is also the most popular

$$\sqrt{[(a_1-b_1)^2 + (a_2-b_2)^2 + \ldots. + (a_n-b_n)^2]}$$

- Chebyshev distance

$$\text{Max} (| a_1 - b_1 | \;, | a_2 - b_2 | \;, \ldots.. | a_n - b_n |)$$

- Manhattan distance

$$| a_1 - b_1 | \quad + \quad | a_2 - b_2 | \quad + \ldots.. \quad | a_n - b_n |$$

K means Clustering

What is Clustering?

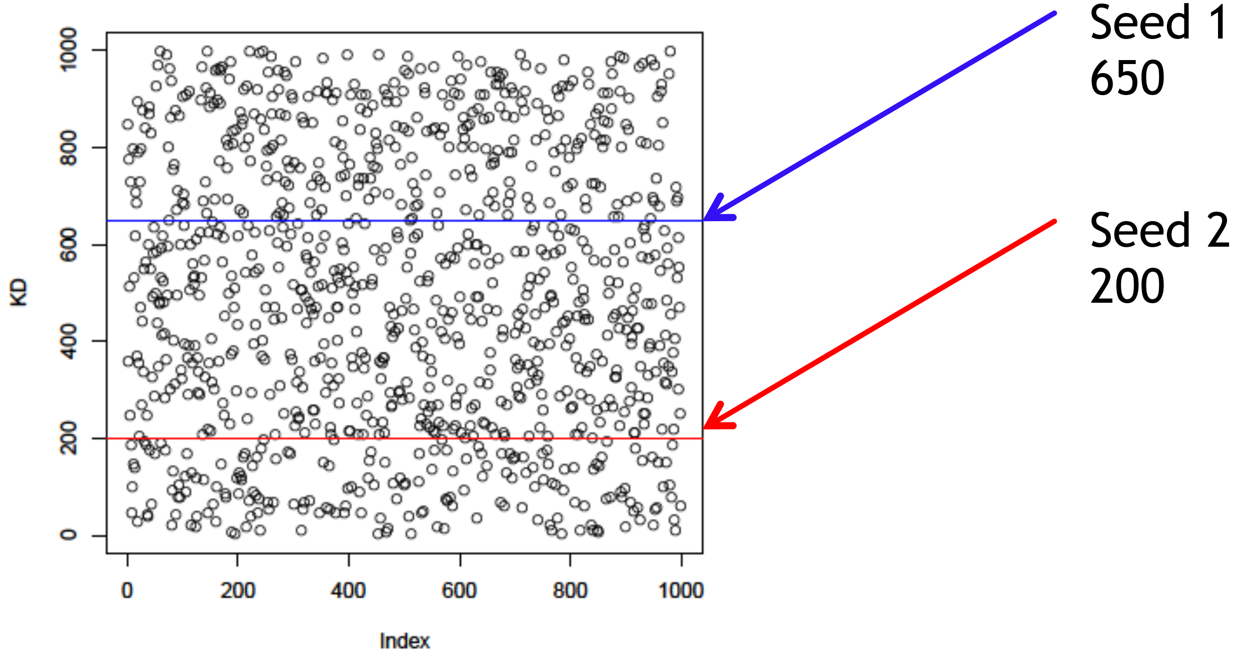
Clustering is dividing data points into homogeneous classes or clusters:

- Points in the same group are as similar as possible
- Points in different group are as dissimilar as possible

When a collection of objects is given, we put objects into group based on similarity.

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solves the well-known clustering problem. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

K means Clustering - Algorithm



Seed 1
650

Seed 2
200

Step 1: Select K . Let us set it at 2.
Step 2: Randomly select initial cluster seeds. An initial cluster seed represents the “mean value” of its cluster.

K means Clustering - Algorithm

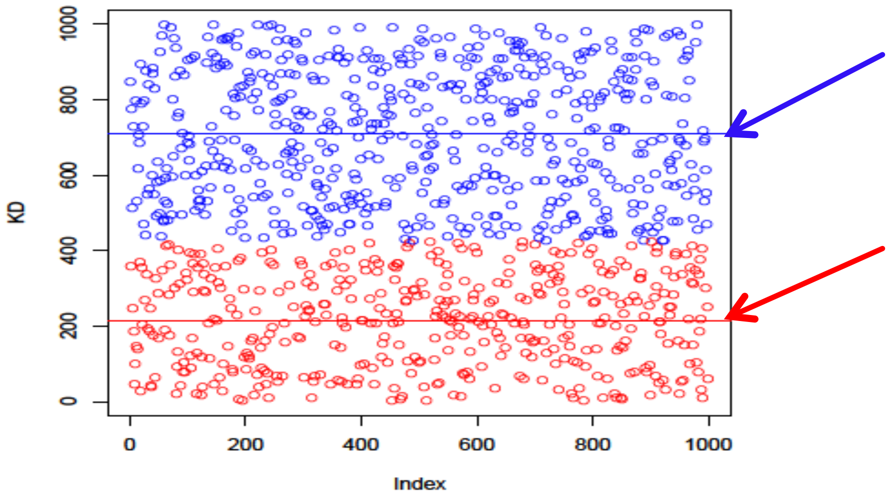
Step 2: calculate distance from each object to each cluster seed.

What type of distance should we use?

- Squared Euclidean distance

Step 3: Assign each object to the closest cluster

Step 4: Compute the new centroid for each cluster



Cluster Seed 1
708.9

Cluster Seed 2
214.2

K means Clustering - Algorithm

Iterate:

- Calculate distance from objects to cluster centroids.
- Assign objects to closest cluster
- Recalculate new centroids

Stop based on convergence criteria

- No change in clusters
- Max iterations

K means Clustering - Issues

Distance measure is squared Euclidean

- Scale should be similar in all dimensions
- Rescale data?
- Not good for nominal data. Why?

Approach tries to minimize the within-cluster sum of squares error (WCSS)

- The overall WCSS is given by: $\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$
- The goal is to find the smallest WCSS
- Implicit assumption that SSE is similar for each group